**PCT** WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C12Q 1/68 — A1

(11) International Publication Number: **WO 96/12822**

(43) International Publication Date: 2 May 1996 (02.05.96)

(21) International Application Number: PCT/SE95/01213

(22) International Filing Date: 17 October 1995 (17.10.95)

(30) Priority Data:
9403612-6 21 October 1994 (21.10.94) SE

(71) Applicant *(for all designated States except US)*: PHARMACIA BIOTECH AB [SE/SE]; S-751 82 Uppsala (SE).

(72) Inventor; and
(75) Inventor/Applicant *(for US only)*: BJÖRKESTEN, Lennart [SE/SE]; Polstjärnevägen 12, S-743 40 Storvreta (SE).

(74) Agents: JOHANSSON, Lars, E. et al.; Bergenstråhle & Lindvall AB, P.O. Box 17704, S-118 93 Stockholm (SE).

(81) Designated States: AU, CA, JP, US, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

**Published**
*With international search report.*
*Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.*

(54) Title: METHOD FOR INDENTIFYING TWO NUCLEIC ACID BASE CODE SEQUENCES

(57) Abstract

In a method and an apparatus for identifying two nucleic acid base code sequences belonging to a given set of known base code sequences and being superposed on each other in an original sequence which comprises base codes as well as ambiguity codes, a master template sequence is constructed from said given set of base code sequences as combination sequences of base codes and ambiguity codes. One or more determinations of the original sequence are made to obtain one or more test sequences which are aligned against said master template sequence in such a manner that the matching between the sequences is optimized. A consensus sequence is determined from the aligned test sequences and is compared with all the combination sequences. A match between one of the combination sequences and the consensus sequence indicates that particular combination sequence corresponds to said two nucleic acide base code sequences to be identified.

*FOR THE PURPOSES OF INFORMATION ONLY*

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

| | | | | | |
|---|---|---|---|---|---|
| **AT** | Austria | **GB** | United Kingdom | **MR** | Mauritania |
| **AU** | Australia | **GE** | Georgia | **MW** | Malawi |
| **BB** | Barbados | **GN** | Guinea | **NE** | Niger |
| **BE** | Belgium | **GR** | Greece | **NL** | Netherlands |
| **BF** | Burkina Faso | **HU** | Hungary | **NO** | Norway |
| **BG** | Bulgaria | **IE** | Ireland | **NZ** | New Zealand |
| **BJ** | Benin | **IT** | Italy | **PL** | Poland |
| **BR** | Brazil | **JP** | Japan | **PT** | Portugal |
| **BY** | Belarus | **KE** | Kenya | **RO** | Romania |
| **CA** | Canada | **KG** | Kyrgystan | **RU** | Russian Federation |
| **CF** | Central African Republic | **KP** | Democratic People's Republic of Korea | **SD** | Sudan |
| **CG** | Congo | | | **SE** | Sweden |
| **CH** | Switzerland | **KR** | Republic of Korea | **SI** | Slovenia |
| **CI** | Côte d'Ivoire | **KZ** | Kazakhstan | **SK** | Slovakia |
| **CM** | Cameroon | **LI** | Liechtenstein | **SN** | Senegal |
| **CN** | China | **LK** | Sri Lanka | **TD** | Chad |
| **CS** | Czechoslovakia | **LU** | Luxembourg | **TG** | Togo |
| **CZ** | Czech Republic | **LV** | Latvia | **TJ** | Tajikistan |
| **DE** | Germany | **MC** | Monaco | **TT** | Trinidad and Tobago |
| **DK** | Denmark | **MD** | Republic of Moldova | **UA** | Ukraine |
| **ES** | Spain | **MG** | Madagascar | **US** | United States of America |
| **FI** | Finland | **ML** | Mali | **UZ** | Uzbekistan |
| **FR** | France | **MN** | Mongolia | **VN** | Viet Nam |
| **GA** | Gabon | | | | |

# METHOD FOR INDENTIFYING TWO NUCLEIC ACID BASE CODE SEQUENCES

## TECHNICAL FIELD

The invention relates to a method and an apparatus for identifying two nucleic acid base code sequences belonging to a given set of known base code sequences and being superposed on each other in an original sequence which comprises base codes as well as ambiquity codes.

## BACKGROUND OF THE INVENTION

Such a method is known from Erik H. Rozemuller et al "Assignment of HLA-DPB alleles by computerized matching based upon sequence data", Human Immunology 37, 207-212 (1993).

According to the known method, a data base containing all known HLA-DPB sequences, makes it possible to analyze heterozygous individuals by combinatorial comparison through all base code sequences and thus identify the one or two base code sequences involved. The HLA-DPB sequences in the data base are selected from published sequences (Marsh S.G.E., Bodmer J.G.; "HLA class II nucleotide sequences", 1992, Tissue Antigens 40:229, 1992).

A disadvantage with the known method is its inability to handle artifacts in terms of inserted or removed base codes in a test sequence.

Moreover, the known method is time consuming and involves a great amount of data.

## BROAD DESCRIPTION OF THE INVENTION

The object of the invention is to bring about a method which is less sensitive to artifacts in non-crucial parts of a test sequence produced by sequencing equipment, when analyzing low quality samples, the artifacts being described in terms of inserted, removed and exchanged base codes and ambiguity codes, and which is less time consuming and involves less data than the known method, as well as an apparatus for carrying that method into effect.

This is attained by a first embodiment of the method according to the invention for identifying two nucleic acid base code sequences belonging to a given set of known base code sequences and being superposed on each other in an original sequence which comprises base codes as well as

ambiquity codes, in that it comprises the steps of
a) constructing a master template sequence from said given set of base code sequences by assigning every conserved position, where the base code is the same all through the set, that particular base code in said master template sequence, and assigning every non-conserved position, where the base code differs through the set, a wild-card code in said master template sequence,
b) extracting from every base code sequence of said given set, the non-conserved positions to obtain non-conserved position subsequences containing only the non-conserved base codes,
c) superposing in pairs all possible combinations of the non-conserved position sequences extracted in step b) to obtain combination sequences of base codes and ambiguity codes,
d) making a determination of the original sequence in order to obtain a test sequence,
e) aligning said test sequence against said master template sequence in such a manner that, accepting gaps in either sequence, the matching between them is optimized, said wild-card coded non-conserved positions in said master template sequence being considered as matching any base code and any ambiguity code in said test sequence,
f) extracting from said test sequence all base codes and ambiguity codes which are aligned with the wild-card codes in said master template sequence, and
g) comparing the base codes and ambiguity codes extracted in step f) with all the combination sequences of base codes and ambiguity codes obtained in step c), a match between one of said combination sequences obtained in step c) and the base codes and ambiguity codes extracted in step f), indicating that that particular combination sequence of base codes and ambiguity codes corresponds to said two nucleic acid base code sequences to be identified.

This is also attained by a second embodiment of the method according to the invention for identifying two nucleic acid base code sequences belonging to a given set of known

base code sequences and being superposed on each other in an original sequence which comprises base codes as well as ambiguity codes, in that it comprises the steps of

a) constructing a master template sequence from said given set of base code sequences by assigning every conserved position, where the base code is the same all through the set, that particular base code in said master template sequence, and assigning every non-conserved position, where the base code differs through the set, a wild-card code in said master template sequence,

b) extracting from every base code sequence of said given set, the non-conserved positions to obtain non-conserved position subsequences containing only the non-conserved base codes,

c) superposing, in pairs, all possible combinations of the non-conserved position sequences extracted in step b) to obtain combination sequences of base codes and ambiguity codes,

d) making one or more determinations of the original sequence in order to obtain one or more test sequences,

e) aligning each of said one or more test sequences against said master template sequence in such a manner that, accepting gaps in either sequence, the matching between the master template and each test sequence is optimized, said wild-card coded non-conserved positions in said master template sequence being considered as matching any base code and any ambiguity code in each test sequence,

f) extracting from each of said test sequences all base codes and ambiguity codes which are aligned with the wild-card codes in said master template sequence,

g) determining, for each non-conserved position, a consensus base code or ambiguity code on the basis of the non-conserved bases extracted from each test sequence by summing up a score for each base code for each non-conserved position and keeping the base code with the highest score, the score being a function of the position of the base code in the respective test sequence as well as of the local quality of the align-

ment between the respective test sequence and said master template sequence, and

h) comparing the consensus base codes and ambiguity codes determined in step g) with all the combination sequences of base codes and ambiguity codes obtained in step c), a match between one of said combination sequences obtained in step c) and the consensus base codes and ambiguity codes determined in step g), indicating that that particular combination sequence of base codes and ambiguity codes corresponds to said two nucleic acid base code sequences to be identified.

A first embodiment of the apparatus according to the invention for identifying two nucleic acid base code sequences belonging to a given set of known base code sequences and being superposed on each other in an original sequence which comprises base codes as well as ambiguity codes, comprises master template sequence constructing means for constructing a master template sequence from said given set of base code sequences by assigning every conserved position, where the base code is the same all through the set, that particular base code in said master template sequence, and assigning every non-conserved position, where the base code differs through the set, a wild-card code in said master template sequence, non-conserved position extracting means for extracting from every base code sequence of said given set, the non-conserved positions to obtain non-conserved position subsequences containing only the non-conserved base codes, superposing means for superposing in pairs all possible combinations of the non-conserved position sequences extracted by said non-conserved position extracting means to obtain combination sequences of base codes and ambiguity codes, original sequence determining means for making a determination of the original sequence in order to obtain a test sequence, aligning means for aligning said test sequence against said master template sequence in such a manner that, accepting gaps in either sequence, the matching between them is optimized, said wild-card coded non-conserved positions in said master template sequence being considered as matching

any base code and any ambiguity code in said test sequence, base code and ambiguity code extracting means for extracting from said test sequence all base codes and ambiguity codes which are aligned with the wild-card codes in said master template sequence, and comparing means for comparing the base codes and ambiguity codes extracted by said base code and ambiguity code extracting means with all the combination sequences of base codes and ambiguity codes obtained by means of said superposing means, a match between one of said combination sequences obtained by means of said superposing means and the base codes and ambiguity codes extracted by said base code and ambiguity code extracting means, indicating that that particular combination sequence of base codes and ambiguity codes corresponds to said two nucleic acid base code sequences to be identified.

A second embodiment of the apparatus according to the invention for identifying two nucleic acid base code sequences belonging to a given set of known base code sequences and being superposed on each other in an original sequence which comprises base codes as well as ambiguity codes, comprises master template sequence constructing means for constructing a master template sequence from said given set of base code sequences by assigning every conserved position, where the base code is the same all through the set, that particular base code in said master template sequence, and assigning every non-conserved position, where the base code differs through the set, a wild-card code in said master template sequence, non-conserved position extracting means for extracting from every base code sequence of said given set, the non-conserved positions to obtain non-conserved position subsequences containing only the non-conserved base codes, superposing means for superposing, in pairs, all possible combinations of the non-conserved position sequences extracted by said non-conserved position extracting means to obtain combination sequences of base codes and ambiguity codes, original sequence determining means for making one or more determinations of the original sequence in order to obtain

one or more test sequences, aligning means for aligning each of said one or more test sequences against said master template sequence in such a manner that, accepting gaps in either sequence, the matching between the master template and each test sequence is optimized, said wild-card coded non-conserved positions in said master template sequence being considered as matching any base code and any ambiguity code in each test sequence, base code and ambiguity code extracting means for extracting from each of said test sequences all base codes and ambiguity codes which are aligned with the wild-card codes in said master template sequence, determining means for determining, for each non-conserved position, a consensus base code or ambiguity code on the basis of the non-conserved bases extracted from each test sequence by summing up a score for each base code for each non-conserved position and keeping the base code with the highest score, the score being a function of the position of the base code in the respective test sequence as well as of the local quality of the alignment between the respective test sequence and said master template sequence, and comparing means for comparing the consensus base codes and ambiguity codes determined by said determining means with all the combination sequences of base codes and ambiguity codes obtained by means of said superposing means, a match between one of said combination sequences obtained by means of said superposing means and the consensus base codes and ambiguity codes determined by said determining means, indicating that that particular combination sequence of base codes and ambiguity codes corresponds to said two nucleic acid base code sequences to be identified.

**DESCRIPTION OF PREFERRED EMBODIMENTS**

In the following description, A, C, G and T stand for adenine, cytosine, guanine and thymine, respectively, while other one-letter codes stand for combinations of nucleotides at the same position as defined by Nomenclature Committee of the International Union of Biochemistry (NC-IUB): Nomen-

clature for incompletely specified bases in nucleic acid sequences. Eur J Biochem 150:1, 1985 as follows:

R = G and A
Y = T and C
W = A and T
S = G and C
M = A and C
K = G and T
B = G and T and C
D = G and A and T
V = G and A and C
H = A and T and C
N = G and A and T and C

In the method according to the invention, one or more determinations of an original sequence are made in order to obtain one or more test sequences. The test sequences are obtained in a manner known per se by means of sequencing equipment, and are to be analyzed in order to identify the two nucleic acid base code sequences which, superposed on each other, make up the original sequence.

To accomplish this, the starting point is a given set of alternative base code sequences (alleles) for a gene in the HLA complex. For this example, the following set of three alternative base code sequences or subtypes could be used:

| | |
|---|---|
| Subtype 1 | ACC GCT GAT CCC TGT CG |
| Subtype 2 | --- --A TG- --- --C G- |
| Subtype 3 | --- --- --- --- --C -- |

According to the nomenclature above, the first subtype is explicitely defined, while merely deviations from the first subtype are indicated for the other two subtypes.

It is to be understood that, in practice, the number of subtypes is very large.

According to the invention a master template sequence is constructed from the above given set of subtypes by assigning

every conserved position, i.e. every position where the base code is the same all through the set, that particular base code in said master template sequence, while every non-conserved position, i.e. every position where the base code differs through the set, is assigned a wild-card code corresponding to $ in said master template sequence.

Applying this to the above given set of just three base code sequences, the master template sequence will be as follows:

ACCGC$$$TCCCTG$$G.

According to the invention, also the non-conserved positions are extracted from every base code sequence in the above given set in order to obtain a corresponding set of non-conserved position subsequences which only contain the non-conserved base codes.

Applying this to the above given set of subtypes, the following three non-conserved position subsequences are obtained:

1. TGATC
2. ATGCG
3. TGACC.

According to the invention, all possible combinations of the above non-conserved position sequences are superposed in pairs in order to obtain combination sequences of base codes and ambiguity codes.

For the above three non-conserved position sequences, the following combination sequences are obtained.

| Combination | |
|---|---|
| 1/1 | TGATC |
| 1/2 | WKRYS |
| 1/3 | TGAYC |
| 2/2 | ATGCG |
| 2/3 | WKRCS |
| 3/3 | TGACC |

In accordance with the invention, a test sequence, obtained as indicated above, is then aligned with the master template sequence in such a manner that, accepting gaps in either sequence, the matching between the test sequence and the master template sequence, is obtimized.

For this alignment, a dynamic programming algorithm described by Sigvard Needleman and C. Wunsch, J. Mol. Biol. 48, 444 (1970), may be used.

This algorithm functions so that all types of alignments between the two sequences are given points. This is accomplished in that different points are awarded e.g. for matching position, mismatching position, inserted or removed characters etc. The alignment that obtains the highest number of points, is kept.

According to the invention, also the wild-card code introduced in accordance with the invention, gives matching points in combination with any character in the other sequence. Thus, the master template sequence will have the function of pointing out non-conserved positions in the test sequence based on the local appearance of the alignment between the sequences. This will function despite different forms of artifacts (inserted, removed and/or exchanged characters) in the conserved regions and without actual knowledge of where the respective test sequence starts.

According to a first embodiment, it is supposed that the below single test sequence has been obtained:

CGGTATCGCWKRTCCCTGCSGGAT.

Aligning the above test sequence and the master template sequence in the above manner would give the following result

```
                              CGGT T                GAT
Test sequence                    A CGCWKRTCCCTGCSG
Master template sequence      ----A CGC$$$TCCCTG$$G---
                                   C
```

According to the invention, all base codes and ambiguity codes which are aligned with the wild-card codes in the master template sequence, are then extracted, which gives the following sequence:

WKRCS.

This extracted sequence of base codes and ambiquity codes is then compared with all the above combination sequences of base codes and ambiguity codes.

A match between one of said combination sequences and the extracted sequence of base codes and ambiguity codes, indicate that that particular combination sequence corresponds to the two nucleic acid base code sequences to be identified.

In this case, the combination 2/3 above corresponds exactly with the extracted sequence, which means that the two nucleic acid base code sequences superposed on each other, in other words, the two HLA alleles for a certain gene present in the sequence obtained from a sample from a human individual, can be identified.

Thus, in the present case, since the subsequences in the combination 2/3 are extracted from subtypes 2 and 3, the test sequence is, in fact, a superposition of subtypes 2 and 3.

According to a second embodiment, it is supposed that the below two test sequences have been obtained:

| | |
|---|---|
| Test sequence I | CGGTATCGCWKRTCCCTGCSGGAT |
| Test sequence II | CGGTACCGTTKRTCCCTGCSGGAT. |

Aligning the above two test sequences and the master template sequence would give the following results:

```
                                CGGT T                  GAT
Test sequence I                    A CGCWKRTCCCTGCSG
Master template sequence        ----A CGC$$$TCCCTG$$G---
                                    C
```

and

```
                                CGGT    T              GAT
Test sequence II                    ACCG TKRTCCCTGCSG
Master template sequence    ----ACCG $$$TCCCTG$$G---
                                      C
```

As in the first embodiment, all base codes and ambiguity codes which are aligned with the wild-card codes in the master template sequence, are then extracted from each test sequence, which gives the following extracted sequences:

WKRCS, and
TKRCS.

According to the invention, when two or more test sequences are obtained, a consensus sequence of base codes and ambiguity codes is then determined from the two or more extracted sequences in the following way:

For each non-conserved position, a score is assigned to all possible code types. For the first position, this gives:

| Code | 1st sequence Score | 2nd sequence Score | Total Score |
|---|---|---|---|
| A | 0 | 0 | 0 |
| C | 0 | 0 | 0 |
| G | 0 | 0 | 0 |
| T | 0 | 0.5-(0.0001*5) | 0.4995 |
| R | 0 | 0 | 0 |
| Y | 0 | 0 | 0 |
| W | 1.0-(0.0001*5) | 0 | 0.9995 |
| S | 0 | 0 | 0 |
| M | 0 | 0 | 0 |
| K | 0 | 0 | 0 |

The code with the highest total score, in this case W=0.9995, is kept for the consensus sequence. The first component, 0.5 and 1.0, respectively, reflects the quality of the local alignment in such a manner that 1.0 means that the

quality of the local alignment is perfect, while 0.5 means that the quality of the local alignment is not perfect, in this case, in view of the mismatch immediately to the left of the position in question. It should be understood that, in this example, 0.5 has been chosen to reflect a mismatch in an adjacent position. The second component, 0.0001*5 in both cases, gives a negative contribution due to the position in the test sequences in such a manner that a position located closer to the beginning of the test sequence gives a smaller negative contribution than a position located further away from the beginning of the test sequence.

The next position is treated in the same way:

| Code | 1st sequence Score | 2nd sequence Score | Total Score |
|---|---|---|---|
| A | 0 | 0 | 0 |
| C | 0 | 0 | 0 |
| G | 0 | 0 | 0 |
| T | 0 | 0 | 0 |
| R | 0 | 0 | 0 |
| Y | 0 | 0 | 0 |
| W | 0 | 0 | 0 |
| S | 0 | 0 | 0 |
| M | 0 | 0 | 0 |
| K | 1.0-(0.0001*6) | 1.0-(0.0001*6) | 1.9988 |

The code with the highest total score, in this case K=1.9988, is kept for the consensus sequence. It should be pointed out that the total score may be used as a quality measure of the position inquestion. Thus, in the above two examples, the quality of K is almost as high as possible.

Treating the rest of the positions in the same manner gives the following final consensus sequence:

WKRCS.

These determined consensus base codes and ambiquity codes are then compared with all the above combination sequences of

base codes and ambiguity codes.

As in the first embodiment, a match between one of said combination sequences and the extracted sequence of base codes and ambiguity codes, indicate that that particular combination sequence corresponds to the two nucleic acid base code sequences to be identified.

Also in this second embodiment, the above combination 2/3 corresponds exactly with the extracted sequence, which means that the two nucleic acid base code sequences superposed on each other, in other words, the two HLA alleles for a certain gene present in the sequence obtained from a sample from a human individual, can be identified.

Thus, also in this case, since the subsequences in the combination 2/3 are extracted from subtypes 2 and 3, the test sequence is, in fact, a superposition of subtypes 2 and 3.

It should be understood that the above second embodiment of the method according to the invention, with two (or more) test sequences, also could be applied to just a single test sequence. In that case, the consensus sequence would, of course, be the same as the test sequence.

A first embodiment of an apparatus according to the invention for identifying two nucleic acid base code sequences belonging to a given set of known base code sequences and being superposed on each other in an original sequence which comprises base codes as well as ambiguity codes, comprises master template sequence constructing means (not shown) for constructing a master template sequence from said given set of base code sequences by assigning every conserved position, where the base code is the same all through the set, that particular base code in said master template sequence, and assigning every non-conserved position, where the base code differs through the set, a wild-card code in said master template sequence, non-conserved position extracting means (not shown) for extracting from every base code sequence of said given set, the non-conserved positions to obtain non-conserved position subsequences containing only the non-conserved base codes, superposing means (not shown) for

superposing in pairs all possible combinations of the non-conserved position sequences extracted by said non-conserved position extracting means to obtain combination sequences of base codes and ambiguity codes, original sequence determining means (not shown) for making a determination of the original sequence in order to obtain a test sequence, aligning means (not shown) for aligning said test sequence against said master template sequence in such a manner that, accepting gaps in either sequence, the matching between them is optimized, said wild-card coded non-conserved positions in said master template sequence being considered as matching any base code and any ambiguity code in said test sequence, base code and ambiguity code extracting means (not shown) for extracting from said test sequence all base codes and ambiguity codes which are aligned with the wild-card codes in said master template sequence, and comparing means (not shown) for comparing the base codes and ambiguity codes extracted by said base code and ambiguity code extracting means with all the combination sequences of base codes and ambiguity codes obtained by means of said superposing means, a match between one of said combination sequences obtained by means of said superposing means and the base codes and ambiguity codes extracted by said base code and ambiguity code extracting means, indicating that that particular combination sequence of base codes and ambiguity codes corresponds to said two nucleic acid base code sequences to be identified.

A second embodiment of an apparatus according to the invention for identifying two nucleic acid base code sequences belonging to a given set of known base code sequences and being superposed on each other in an original sequence which comprises base codes as well as ambiguity codes, comprises master template sequence constructing means (not shown) for constructing a master template sequence from said given set of base code sequences by assigning every conserved position, where the base code is the same all through the set, that particular base code in said master template sequence, and

assigning every non-conserved position, where the base code differs through the set, a wild-card code in said master template sequence, non-conserved position extracting means (not shown) for extracting from every base code sequence of said given set, the non-conserved positions to obtain non-conserved position subsequences containing only the non-conserved base codes, superposing means (not shown) for superposing, in pairs, all possible combinations of the non-conserved position sequences extracted by said non-conserved position extracting means to obtain combination sequences of base codes and ambiguity codes, original sequence determining means (not shown) for making one or more determinations of the original sequence in order to obtain one or more test sequences, aligning means (not shown) for aligning each of said one or more test sequences against said master template sequence in such a manner that, accepting gaps in either sequence, the matching between the master template and each test sequence is optimized, said wild-card coded non-conserved positions in said master template sequence being considered as matching any base code and any ambiguity code in each test sequence, base code and ambiguity code extracting means (not shown) for extracting from each of said test sequences all base codes and ambiguity codes which are aligned with the wild-card codes in said master template sequence, determining means (not shown) for determining, for each non-conserved position, a consensus base code or ambiguity code on the basis of the non-conserved bases extracted from each test sequence by summing up a score for each base code for each non-conserved position and keeping the base code with the highest score, the score being a function of the position of the base code in the respective test sequence as well as of the local quality of the alignment between the respective test sequence and said master template sequence, and comparing means (not sshown) for comparing the consensus base codes and ambiguity codes determined by said determining means with all the combination sequences of base codes and ambiguity codes obtained by means of said superposing means,

a match between one of said combination sequences obtained by means of said superposing means and the consensus base codes and ambiguity codes determined by said determining means, indicating that that particular combination sequence of base codes and ambiguity codes corresponds to said two nucleic acid base code sequences to be identified.

The apparatuses according to the invention are preferably implemented in computer software.

**CLAIMS**

1. A method for identifying two nucleic acid base code sequences belonging to a given set of known base code sequences and being superposed on each other in an original sequence which comprises base codes as well as ambiguity codes, **characterized by** the steps of

a) constructing a master template sequence from said given set of base code sequences by assigning every conserved position, where the base code is the same all through the set, that particular base code in said master template sequence, and assigning every non-conserved position, where the base code differs through the set, a wild-card code in said master template sequence,

b) extracting from every base code sequence of said given set, the non-conserved positions to obtain non-conserved position subsequences containing only the non-conserved base codes,

c) superposing, in pairs, all possible combinations of the non-conserved position sequences extracted in step b) to obtain combination sequences of base codes and ambiguity codes,

d) making a determination of the original sequence in order to obtain a test sequence,

e) aligning said test sequence against said master template sequence in such a manner that, accepting gaps in either sequence, the matching between them is optimized, said wild-card coded non-conserved positions in said master template sequence being considered as matching any base code and any ambiguity code in said test sequence,

f) extracting from said test sequence all base codes and ambiguity codes which are aligned with the wild-card codes in said master template sequence, and

g) comparing the base codes and ambiguity codes extracted in step f) with all the combination sequences of base codes and ambiguity codes obtained in step c), a match between one of said combination sequences obtained in step c) and the base

codes and ambiguity codes extracted in step f), indicating that that particular combination sequence of base codes and ambiguity codes corresponds to said two nucleic acid base code sequences to be identified.

2. A method for identifying two nucleic acid base code sequences belonging to a given set of known base code sequences and being superposed on each other in an original sequence which comprises base codes as well as ambiguity codes, **characterized by** the steps of

a) constructing a master template sequence from said given set of base code sequences by assigning every conserved position, where the base code is the same all through the set, that particular base code in said master template sequence, and assigning every non-conserved position, where the base code differs through the set, a wild-card code in said master template sequence,

b) extracting from every base code sequence of said given set, the non-conserved positions to obtain non-conserved position subsequences containing only the non-conserved base codes,

c) superposing, in pairs, all possible combinations of the non-conserved position sequences extracted in step b) to obtain combination sequences of base codes and ambiguity codes,

d) making one or more determinations of the original sequence in order to obtain one or more test sequences,

e) aligning each of said one or more test sequences against said master template sequence in such a manner that, accepting gaps in either sequence, the matching between the master template and each test sequence is optimized, said wild-card coded non-conserved positions in said master template sequence being considered as matching any base code and any ambiguity code in each test sequence,

f) extracting from each of said test sequences all base codes and ambiguity codes which are aligned with the wild-card codes in said master template sequence,

g) determining, for each non-conserved position, a consensus base code or ambiguity code on the basis of the non-conserved bases extracted from each test sequence by summing up a score for each base code for each non-conserved position and keeping the base code with the highest score, the score being a function of the position of the base code in the respective test sequence as well as of the local quality of the alignment between the respective test sequence and said master template sequence, and

h) comparing the consensus base codes and ambiguity codes determined in step g) with all the combination sequences of base codes and ambiguity codes obtained in step c), a match between one of said combination sequences obtained in step c) and the consensus base codes and ambiguity codes determined in step g), indicating that that particular combination sequence of base codes and ambiguity codes corresponds to said two nucleic acid base code sequences to be identified.

3. A method of genetic analysis, comprising the steps of

(i) subjecting a test sample to a sequencing procedure to obtain two superposed base code sequences representing the alleles present for a specific gene, and

(ii) identifying the base code sequences by the method according to claim 1 or 2.

4. Use of the method according to claim 1 or 2 for HLA typing.

5. An apparatus for identifying two nucleic acid base code sequences belonging to a given set of known base code sequences and being superposed on each other in an original sequence which comprises base codes as well as ambiguity codes, **characterized in that** it comprises

- master template sequence constructing means for constructing a master template sequence from said given set of base code sequences by assigning every conserved position, where the base code is the same all through the set, that particu-

lar base code in said master template sequence, and assigning every non-conserved position, where the base code differs through the set, a wild-card code in said master template sequence,

- non-conserved position extracting means for extracting from every base code sequence of said given set, the non-conserved positions to obtain non-conserved position subsequences containing only the non-conserved base codes,
- superposing means for superposing, in pairs, all possible combinations of the non-conserved position sequences extracted by said non-conserved position extracting means to obtain combination sequences of base codes and ambiguity codes,
- original sequence determining means for making a determination of the original sequence in order to obtain a test sequence,
- aligning means for aligning said test sequence against said master template sequence in such a manner that, accepting gaps in either sequence, the matching between them is optimized, said wild-card coded non-conserved positions in said master template sequence being considered as matching any base code and any ambiguity code in said test sequence,
- base code and ambiguity code extracting means for extracting from said test sequence all base codes and ambiguity codes which are aligned with the wild-card codes in said master template sequence, and
- comparing means for comparing the base codes and ambiguity codes extracted by said base code and ambiguity code extracting means with all the combination sequences of base codes and ambiguity codes obtained by means of said superposing means, a match between one of said combination sequences obtained by means of said superposing means and the base codes and ambiguity codes extracted by said base code and ambiguity code extracting means, indicating that that particular combination sequence of base codes and ambiguity codes corresponds to said two nucleic acid base code sequences to be identified.

6. An apparatus for identifying two nucleic acid base code sequences belonging to a given set of known base code sequences and being superposed on each other in an original sequence which comprises base codes as well as ambiguity codes, **characterized in** that it comprises

- master template sequence constructing means for constructing a master template sequence from said given set of base code sequences by assigning every conserved position, where the base code is the same all through the set, that particular base code in said master template sequence, and assigning every non-conserved position, where the base code differs through the set, a wild-card code in said master template sequence,
- non-conserved position extracting means for extracting from every base code sequence of said given set, the non-conserved positions to obtain non-conserved position subsequences containing only the non-conserved base codes,
- superposing means for superposing, in pairs, all possible combinations of the non-conserved position sequences extracted by said non-conserved position extracting means to obtain combination sequences of base codes and ambiguity codes,
- original sequence determining means for making one or more determinations of the original sequence in order to obtain one or more test sequences,
- aligning means for aligning each of said one or more test sequences against said master template sequence in such a manner that, accepting gaps in either sequence, the matching between the master template and each test sequence is optimized, said wild-card coded non-conserved positions in said master template sequence being considered as matching any base code and any ambiguity code in each test sequence,
- base code and ambiguity code extracting means for extracting from each of said test sequences all base codes and ambiguity codes which are aligned with the wild-card codes in said master template sequence,
- determining means for determining, for each non-conserved position, a consensus base code or ambiguity code on the

basis of the non-conserved bases extracted from each test sequence by summing up a score for each base code for each non-conserved position and keeping the base code with the highest score, the score being a function of the position of the base code in the respective test sequence as well as of the local quality of the alignment between the respective test sequence and said master template sequence, and

- comparing means for comparing the consensus base codes and ambiguity codes determined by said determining means with all the combination sequences of base codes and ambiguity codes obtained by means of said superposing means, a match between one of said combination sequences obtained by means of said superposing means and the consensus base codes and ambiguity codes determined by said determining means, indicating that that particular combination sequence of base codes and ambiguity codes corresponds to said two nucleic acid base code sequences to be identified.

# INTERNATIONAL SEARCH REPORT

International application No.
PCT/SE 95/01213

## A. CLASSIFICATION OF SUBJECT MATTER

IPC6: C12Q 1/68
According to International Patent Classification (IPC) or to both national classification and IPC

## B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC6: C12Q, H03M

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

SE,DK,FI,NO classes as above

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CA, BIOSIS, MEDLINE, SCISEARCH, PATENT CITATION INDEX

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| X | Human Immunology, Volume 37, 1993, Erik H. Rozemuller et al, "Assignment of HLA-DPB Alleles by Computerized Matching Based Upon Sequence Data" page 207 - page 212 -- | 1-6 |
| X | J.Mol.Biol., Volume 221, 1991, B. Edwin Blaisdell et al, "An Efficient Algorithm for Identifying Matches with Errors in Multiple Long Molecular Sequences", page 1367 - page 1378, the whole document especially p 1370 c1, line 29 - C2 line 43 -- -------- | 1-6 |

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" erlier document but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance: the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance: the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

| Date of the actual completion of the international search | Date of mailing of the international search report |
|---|---|
| 20 February 1996 | 27. 02- 1996 |
| Name and mailing address of the ISA/<br>Swedish Patent Office<br>Box 5055, S-102 42 STOCKHOLM<br>Facsimile No. +46 8 666 02 86 | Authorized officer<br>Patrick Andersson<br>Telephone No. +46 8 782 25 00 |